# Early Syntactic Bootstrapping in an Incremental Memory-Limited Word Learner

**Sepideh Sadeghi and Matthias Scheutz**
Computer Science Department
Tufts University, Medford MA, USA
{sepideh.sadeghi,mscheutz}@tufts.edu

## Abstract

It has been suggested that early human word learning occurs across learning situations and is bootstrapped by syntactic regularities such as word order. Simulation results from ideal learners and models assuming prior access to structured syntactic and semantic representations suggest that it is possible to jointly acquire word order and meanings and that learning is improved as each language capability bootstraps the other. We first present a probabilistic framework for early syntactic bootstrapping in the absence of advanced structured representations, then we use our framework to study the utility of joint acquisition of word order and word referent and its onset, in a memory-limited incremental model. Comparing learning results in the presence and absence of joint acquisition of word order in different ambiguous contexts, improvement in word order results showed an immediate onset, starting in early trials while being affected by context ambiguity. Improvement in word learning results on the other hand, was hindered in early trials where the acquired word order was imperfect, while being facilitated by word order learning in future trials as the acquired word order improved. Furthermore, our results showed that joint acquisition of word order and word referent facilitates one-shot learning of new words as well as inferring intentions of the speaker in ambiguous contexts.

## Introduction

A hallmark of human word learning is the integration of *cross-situational information* even though this information is not always reliable as inconsistencies in the word-referent co-occurrence (e.g., when the referent is absent in a scene or when distracting referents are present) inject noise into cross-situational information. It has been suggested that bootstrapping cross-situational word learning with the learner's belief about the referential intentions of the speaker (Frank, Goodman, and Tenenbaum 2009) as well as bootstrapping it with learner's belief about the syntactic regularities of language (Yu 2006; Maurits, Perfors, and Navarro 2009; Alishahi and Fazly 2010; Alishahi and Chrupała 2012; Abend et al. 2017) allow for disambiguation and should thus improve word learning. Maurits, Perfors, and Navarro (2009) bootstrap word learning with the acquired knowledge of word order in an ideal learner although their model cannot handle variable-length utterances. Alishahi and Fazly (2010) showed that knowledge of lexical categories improves word learning results. Yu and colleagues went further and showed that imperfect knowledge of syntactic regularities learned in parallel with word meaning improves word learning results (Yu 2006; Alishahi and Chrupała 2012). Abend et al. (2017) recently went even further and proposed a truly joint learner in which the learned meanings are used to refine the syntactic knowledge, an aspect missing in the previously proposed joint learners. However, all of these models studied the problem of joint acquisition in the context of ideal learners, ignoring the possibility that the memory and computational limitations of a learner (e.g., an embodied robot) can turn the positive effect of joint acquisition into a negative effect by injecting noise to it. Furthermore, they usually assume prior access to syntactic concepts such as "subjecthood", lexical categories or more advanced syntactic knowledge such as syntactic parses of the input sentences. Abend et al. (2017) use pairs of utterance and utterance semantic representation as input, and their semantic representations mirror the syntactic parse representations of the utterance using lambda calculus. For example, the semantic representation of the sentence "you get a fly" would be $\{\lambda ev.v|get(pro|you, det|a(e, n|fly(e)), ev\}$.

This work explores the possibility of learning word order before syntactic concepts such as subject, object, or lexical categories or syntactic parse representations are available to the learner. It also examines the utility of the acquired word order in a joint learner where word order knowledge constrains word learning (syntactic bootstrapping) and vice versa. We propose that the transitional probabilities of the thematic roles (in the order of their appearance in the utterance) of the referential words (words with action or event participant referents) are an invaluable source of information for learning word order and that they can provide an initial understanding of the notion of word order in early stages of language acquisition in the absence of advanced syntactic concepts or representations. We utilize an incremental and memory-limited learning algorithm as opposed to batch learning algorithms, as we are interested in online learning in embodied agents with computational limitations. Our model adds the notion of syntax to the word learning generative story in (Sadeghi, Scheutz, and Krause 2017; Frank, Goodman, and Tenenbaum 2009) but departs from

previous attempts (Sadeghi and Scheutz 2017) in its design (graphical model) and the information source used to learn about word order, which collectively enable the model to not only learn word order and word referent jointly, but to also handle real-world variable-length utterances, an important aspect missing in (Sadeghi and Scheutz 2017).

## Model Overview

We assume that the learner is capable of object and action categorization prior to word learning, but we do not assume any prior syntactic knowledge. Our model seeks to identify the *referential words* (words with action or object referents in the scene), infer their correct referents and store the correct word-referent mappings in the *lexicon*, which is a many-to-many mapping between words and referents. The input to the model consists of *word learning situations* (or *trials*), each of which is comprised of a scene description paired with an utterance. The scene description consists of a list of semantic predicates corresponding to the unique events happening in the scene and the utterance as an ordered set of words. Scene and utterance may be empty lists. The events listed in the scene description are not necessarily the ones talked about by the speaker and the model relies on what it already knows about the words and their referents from its *lexicon* to identify the referential intentions of the speaker. We use the term *referential intention* in this paper to refer to the event that is listed in the scene description and that the speaker is talking about.

### Event Representation

A scene description consists of a list of *semantic predicates* (or "event representations"). Each semantic predicate corresponds to a unique event occurring in the scene and is represented as a list of event participant and thematic role pairs. For example, the event of "mom gave Lily a doll" would be represented as {(agent MOM) (action GIVE) (patient1 LILY) (patient2 DOLL)} where the action "GIVE" glues several event participants to each other to provide a certain degree of detail about who performed the action and how, when, and where (see Table 3 for more examples).

### Early Syntactic Bootstrapping

Here, we propose that identifying the referential words of the utterance is an important step towards learning structural rules of language and syntactic bootstrapping. We use the term "early syntactic bootstrapping" to refer to the kind of syntactic bootstrapping that can occur in early stages of acquisition when the concepts of NP, VP, adverbs, determiners and other NP/VP modifiers are unknown to the learner. Our account of early syntactic bootstrapping assumes that the concepts of concrete objects and actions are available to the learner and that the learner begins by learning the label of objects (event participants) and actions (events). We propose that tracking the relative order by which event participants are referred to in sentences allows for learning the transitional or bigram probabilities of the thematic roles, which, in turn, provides an initial notion of word order. For example, observing that the word referring to the action performer is always or often the first referential word in the sentence and that the action word is most likely to follow the action performer in the sentence and not vice versa, facilitates learning the "SV" part of the prominent "SVO" English word order. Seeking to learn the labels of the observed objects and actions allows the learner to filter out other words and notice the relative order of referential words in the sentence. This facilitates discovering that English sentences are most likely to start with a referential word for "action performer", which is most likely to be followed by a referential word for "action", which in turn is most likely to be followed by a referential word for "patient".

### Word Order Representation

The notion of word order $\Theta$ consists of $n$ multinomial probability distributions $\theta_{role_i}$ corresponding to $n$ thematic roles $role_i$ known by the learner, where $\theta_{role_i}$ refers to $P(.|role_i)$ defined over all thematic roles. Therefore, $\theta_{role_i}$ consists of $n$ bigram transitional probabilities $\pi_{(role_j|role_i)}$ for each thematic role $role_j$. The model starts with uniform probability distributions over all thematic roles for each $\theta_{role_i}$.

The number and notion of the thematic roles known to the learner probably should evolve as more situations and events are encountered, but here we assume a fixed number of known roles during word learning. Note that this notion of word order does not depend on any prior syntactic knowledge such as concepts of *subject*, *object*, or *verb*. Also note that word order acquisition in our model is built on two assumptions. First, the learner assumes that all utterances made by the speaker follow a consistent word order. Second, the learner tracks the relative order of the referential words and the thematic roles of their referents.

## Models

Here, we present two word learning models. The first model is M-WO which seeks to learn word order and word referent jointly. M-WO allows for the acquired word order information to constrain the acquisition of words meanings and vice versa. M-B on the other hand, only seeks to learn the referent of words and is used as the baseline model to examine the utility of joint acquisition of word order and word referent. Fig. 1 represents the design of the M-WO (with $\Theta$) and M-B models (without $\Theta$), along with their word learning variables and their probabilistic dependencies.

The learner assumes that in each situation, the speaker uses the generative process illustrated in Fig. 1 to produce an utterance ($W_s$) corresponding to the scene ($E_s$). The goal of the learner is to reverse this generative process and infer the lexicon ($L$) and the word order ($\Theta$) used by the speaker. $\gamma$ is a model parameter, capturing the probability that any word in the utterance may have a referent in the scene. $P_R$ is the probability by which a particular word in the utterance may be uniformly chosen from the lexicon to refer to a particular referent in $I_s$. The probability of non-referential use of words ($P_{NR}$), is set to $\kappa$ for words in the model lexicon (to penalize the non-referential use of referential words), and is set to 1 for other words. We use two notions of lexicon: (1) (full) *lexicon* (global hypothesis) and (2)

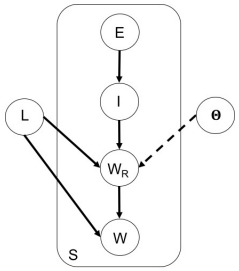



Figure 1: Graphical model describing the generation of utterance ($W$) and its referential words ($W_R$) given the intention ($I$), lexicon ($L$) and word order ($\Theta$) in the context of one situation ($S$). Note that the plate notation signifies that if multiple situations ($C$) were accessible in each trial (e.g., if the model could remember more than one situation at a time), the same relations would hold for all $S$ in $C$. The intention ($I$) is drawn uniformly from the events ($E$) present in the scene in each situation ($S$). $W_R$ is a subset of $W$ which includes only words that refer to an object or action in $I$.

*mini-lexicon* (local hypothesis). "Mini-lexicon" refers to the context-appropriate portions of the full lexicon where context refers to the current situation. The model infers a mini-lexicon in each situation and the lexicon is built through the incremental aggregation of the mini-lexica.

## M-WO

In each situation the model infers the context-appropriate part of the speaker's lexicon ($L$) which has been used to generate the current utterance ($W$) in accordance with $I$ and $\Theta$, where context refers to the current situation. In doing so, the model tries to maximize the joint posterior probability of mini-lexica and word order hypotheses according to the Bayes equation and the probability distribution that the model defines over unobserved lexica ($L$), word order ($\Theta$), and the available context-appropriate evidence ($C$). In this paper, $C$ only includes the current situation, but it can incorporate multiple situations.

$$P(L, \Theta|C) \propto P(C|L, \Theta)P(L)P(\Theta) \tag{1}$$

We assume that $P(\Theta) \propto 1$ for different word orders, and $P(L) \propto e^{-\alpha \cdot |L|}$ serving as a soft mutual exclusivity constraint to produce a preference for one-to-one mappings in the mini-lexica inferred in each situation. Given the probabilistic structure of the model and the fact that speakers' referential intentions and referential words of the utterance are not observable, we marginalize over all possible intentions and possible set of referential words in each situation and rewrite the likelihood term $P(C|L, \Theta)$ as:

$$P(C|L, \Theta) = \prod_{s \in C} \sum_{I_s \subseteq E_s} \sum_{W_R \subseteq W_s} P(W_s|I_s, W_R, L, \Theta) \cdot P(I_s|E_s) \tag{2}$$

Assuming that $P(I_s|E_s) \propto 1$ and that the words of the utterance are generated independently, we can rewrite the term $P(W_s|I_s, W_R, L, \Theta)$ as:

$$P(W_s|I_s, W_R, L, \Theta) = P(W_R|I_s, L, \Theta) \cdot P(W_{NR}|L) \tag{3}$$

where $W_{NR} = \{W_s - W_R\}$. $P(W_R|I_s, L, \Theta)$ then computes the probability of generating the referential words taking into account the relative order of their appearance in the utterance, given $I_s$, $L$, $\Theta$.

$$P(W_R|I_s, L, \Theta) = \prod_{w_j \in W_R} \gamma \sum_{x_i \in I_s} \frac{1}{|I_s|} P_R(w_j|x_i, L) \cdot P(w_{j-1}|ref(w_j) = x_i, I_s, L, \Theta) \tag{4}$$

If $w_j$ is the first referential word in the utterance ($j = 0$), then:

$$P(w_{j-1}|ref(w_j) = x_i, I_s, L, \Theta) = \pi_{(role(x_i)|none)} \tag{5}$$

where $\pi_{(role(x_i)|none)}$ captures the probability of the first referential word in the utterance taking on the thematic role of $x_i$; otherwise:

$$P(w_{j-1}|ref(w_j) = x_i, I_s, L, \Theta) = \sum_{y_k \in I_s} \frac{1}{|I_s|} P_R(w_{j-1}|y_k, L)\pi_{(role(x_i)|role(y_k))} \tag{6}$$

$P(W_{NR}|L)$, on the other hand, is the probability of generating the non-referential words of the utterance given the lexicon $L$.

$$P(W_{NR}|L) = \prod_{w_k \in W_{NR}} (1 - \gamma) P_{NR}(w_k|L) \tag{7}$$

## M-B

The goal of the M-B model is to find the the MAP ("maximum a posteriori") lexicon according to $P(L|C) \propto P(C|L)P(L)$, where the likelihood term $P(C|L)$ can be rewritten as:

$$P(C|L) = \prod_{s \in C} \sum_{I_s \subseteq E_s} P(W_s|I_s, W_R, L) P(I_s|E_s) \tag{8}$$

and we can rewrite the term $P(W_s|I_s, W_R, L)$ as:

$$P(W_s|I_s, W_R, L) = P(W_R|I_s, L) \cdot P(W_{NR}|L) \tag{9}$$

$$P(W_R|I_s, L) = \prod_{w_j \in W_R} \gamma \sum_{x_i \in I_s} \frac{1}{|I_s|} P_R(w_j|x_i, L) \cdot P(w_{j-1}|ref(w_j) = x_i, I_s, L) \tag{10}$$

If $w_j$ is the first referential word in the utterance ($j = 0$) then:

$$P(w_{j-1}|ref(w_j) = x_i, I_s, L) \propto 1 \tag{11}$$

Otherwise:

$$P(w_{j-1}|ref(w_j) = x_i, I_s, L) \propto \sum_{y_k \in I_s} \frac{1}{|I_s|} P_R(w_{j-1}|y_k, L) \tag{12}$$

$P(W_{NR}|L)$ is computed the same way as it is computed in M-WO.

## Incremental Word Learning

We use a variation of the incremental and memory-limited learning algorithm proposed in (Sadeghi and Scheutz 2017) which accounts for the real-world constraints faced by word learners (e.g., infants or robots). This algorithm, to the best of our knowledge, is the only one which satisfies the set of constraints we define on realistic incremental learning. These constraints include: (1) seeing each situation only once with no iteration over data, (2) using only the acquired knowledge and the current observation for hypothesis generation and evaluation, and (3) maintaining a single global hypothesis across different situations motivated by recent findings in (Medina et al. 2011). Our constraints exclude the use of many proposed incremental algorithms in the literature (Liang, Jordan, and Klein 2009; Pearl, Goldwater, and Steyvers 2010; Börschinger and Johnson 2011; 2012). In the rest of this section we first give a high-level overview of the incremental learning algorithm, while highlighting the differences between our version and the original version proposed in (Sadeghi and Scheutz 2017). We refer the reader to (Sadeghi and Scheutz 2017) for the details shared between the two algorithms.

Analogous to (Sadeghi and Scheutz 2017), our learning algorithm is composed of two major components: (1) inferring the MAP mini-lexicon in each situation, and (2) integrating the new mini-lexicon in the previous lexicon, using conflict resolution on conflicting mappings. Inferring the MAP mini-lexicon, subsequently has two components: generating mini-lexicon proposals and scoring the generated mini-lexica. Scoring is performed by computing the relative posterior probability of the mini-lexicon proposals based on the Bayes equations described earlier for M-WO and M-B. Our learning algorithm allows for departing from ideal learners due to its limited memory of past observations (limited access to evidence and co-occurrence statistics) as well as its limited application of Bayesian inference for hypothesis evaluation (scoring). Bayesian inference is only used locally, for hypothesis evaluation in the context of single situations. Prior to Bayesian inference, the model needs to generate several mini-lexica (groups of word-referent mappings). The information used for mini-lexicon generation in each situation, consists of the current situation and the context-appropriate word-referent mappings in the memory (i.e., word-referent pairs stored in the lexicon as opposed to all word-referent pairs encountered so far). Generating mini-lexicon proposals is guided by semi-stochastic search techniques analogous to (Sadeghi and Scheutz 2017). The evidence used for scoring (Bayesian inference) in each situation consists of only the current situation. Different mini-lexica are generated during hypothesis generation and evaluated during hypothesis evaluation. Then, the mappings in the best local hypothesis (mini-lexicon) are added to the global hypothesis (lexicon), performing conflict resolution using the uncertainty associated with conflicting mappings. Specifically, the co-occurrence statistics are used as a measure of the model's uncertainty in the correctness of conflicting mappings. The model maintains only one global hypothesis in all situations and it only makes local revisions to the context appropriate parts of the global hypothesis in light of the current evidence.

Our algorithm departs from the original version in: (1) how it handles conflict resolution when integrating the newly inferred mini-lexicon in the previous lexicon, and (2) the evidence it uses for hypothesis testing in each situation. Our model uses only the current situation as the evidence for hypothesis testing in each situation. This is to avoid the complications that arise from not knowing the roles fulfilled by a particular referent in the lexicon. In (Sadeghi and Scheutz 2017), during conflict resolution, alternative mappings compete with each other and only mappings with the highest co-occurrence statistics are allowed to be included in the lexicon. This strict mutual exclusivity constrain not only inhibits learning of alternative word-referent mappings (e.g., "dog","dogie", and "puppy" all refer to the same concept "DOG"), but also destabilizes the learning results. Our model on the other hand, modulates their strict mutual exclusivity constraint by allowing the addition of alternative mappings for each object, if their co-occurrence statistics fall within a certain range of the co-occurrence statistics of the best existing mapping for that object in memory (lexicon and the current situation).

## Incremental Word Order Learning

Word order learning consists of updating $\theta_{role_i} \in \Theta$ based on the current best lexicon, and the current role bigram counts after processing in each situation. We use a symmetric Dirichlet distribution (with parameter $\beta$) as the conjugate prior for each multinomial distribution $\theta_{role_i}$. Large values of $\beta$ represent a strong prior bias toward nonsparsity and small values represent a strong bias toward sparsity of $\theta_{role_i}$ (multinomial distributions). The value of each $\pi_{(role_j|role_i)}$ at initialization is $\beta/n\beta$), where $n$ is the total number of roles. As the model receives more input incrementally, it updates each $\pi_{(role_j|role_i)} \in \Theta$:

$$\pi_{(role_j|role_i)} = \frac{Count(role_j|role_i) + \beta}{\sum_k Count(role_k|role_i) + n\beta} \tag{13}$$

## Evaluation Data

We evaluate M-WO and M-B in different ambiguous contexts using the datasets described in Table 1 (each dataset consists of 500 trials). These datasets differ from each other in the source and level of their ambiguity. D1 is the least ambiguous, D2 is linguistically more ambiguous than D1, and D3 is visually more ambiguous than D1. D1 and D3 use similar utterances. D1 and D2 use similar scenes. D2 utterances are generated by random addition of two non-referential words (adjective or determiner) from our data-generation lexicon to half of the D1 utterances. D3 scenes include the events used in D1 scenes and an additional event corresponding to an alternative description of the same event for verbs that allow such possibility. This type of ambiguity occurs when a particular scene can be described using (1) both transitive and intransitive verbs such as "drop" and "fall" in "dad dropped the box" and "the box fell", and (2) two different verbs describing the event from the perspective of different event participants,

such as "give" and "take" in "dad gave mom the key" and "mom took the key from dad". This type of ambiguity adds distracting events to the scene description that have high degrees of semantic overlap with the target event and, therefore, are harder to disambiguate compared to distracting events that are added at random or from the nearby utterances in the data (Fazly, Alishahi, and Stevenson 2010; Abend et al. 2017). For instance, the above examples for "take" and "give" share the same objects except that the role of "mom" and "dad" are different in these two sentences. We used a probabilistic generative process to automatically create 500 utterances for D1, with 10 verbs={falls,drops,pushes,pulls,takes,gives,eats,feeds, drinks,reads} and 20 objects. Our verbs were selected from the 13 most frequent verbs in the Brown corpus (Brown 1973; Brown and Bellugi 1964) of the CHILDES database (MacWhinney 2000) and some additional verbs which allow for alternative event descriptions. For each verb in our lexicon, we listed a set of possible frames and used a uniform distribution over them for utterance generation. Then we selected 20 objects from the Brown corpus which were most likely to be used in our frames and added them to our data-generation lexicon. Our data-generation lexicon also includes five prepositions, ten adjectives and three determiners. Overall, we used 48 frames (with SVO word order), a subset of which is depicted in Table 2. After generating the utterances of D1, D2, and D3, we manually generated the corresponding event representations for each utterance while adding distracting events to the scenes of D3. In addition to the big datasets (D1, D2 and D3), we use the following three small datasets to evaluate the utility of joint acquisition of word order and word referent in facilitating "one-shot" learning: dataset D4 with 10 trials, new referential words in its utterances and new referents in its scenes; dataset D5 with five trials, one new referential word and one new non-referential word per utterance and one new referent per scene; and dataset D6 with five trials, one or two new referential words per utterance and two or three events in each scene. We varied the source of ambiguity in these datasets to assess the salience of utility of joint acquisition in facilitating one-shot learning. Example datapoints from D4, D5, and D6 are demonstrated in Table 3.

Table 1: Sources of ambiguity in evaluation data.

| Data | Distracting Events | Non-Ref Words | Prepositions |
|---|---|---|---|
| D1 | No | No | 5 |
| D2 | No | 1 per utterance on average | 5 |
| D3 | 1 for verbs in {drop,feed,take,give} | No | 5 |

## Results

All results are averaged over five runs. The choice of best parameter values to maximize the word learning results depends on the input dataset. We ran M-B on D1, using different parameter values to find a good set of parameters which are used in all of our simulations with both M-B and M-WO: $\gamma = 0.9$, $\alpha = 10$, $\kappa = 0.1$, and $\beta = 1$ (used in M-WO only).

Table 2: Example verb frames for "eat". Except for terminals, the rest of the variables are place-holders for a set of other terminals, variables or a combination of both.

| ID | Frame |
|---|---|
| 1 | Modifier N-animate (terminal 'eat) |
| 2 | Modifier N-animate (terminal 'eat) Modifier N-edible |
| 3 | Modifier N-animate (terminal 'eat) Modifier N-edible (terminal 'in) LOC |

Table 3: Example datapoints from D4, D5, D6 (test datasets) which correspondingly contain 10, 5, and 5 situations. These datasets were generated manually using 2 or 3 thematic roles including "ag", "ac", and "th" which correspondingly refer to "agent", "action", and "theme". Each scene is a list of unique events and each unique event is a list of (role REFERENT) pairs enclosed in $\{\}$.

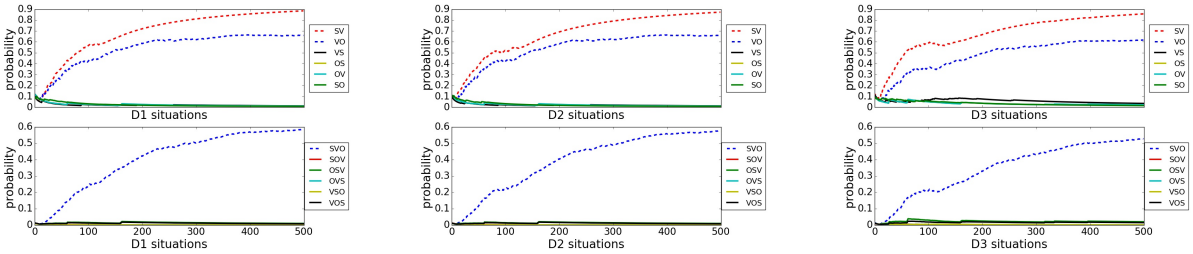
| Data | Utterance | Scene |
|---|---|---|
| D4 | sister plays piano | {(ag SISTER) (ac PLAY) (th PIANO)} |
| D5 | tall girl eats chocolate | {(ag GIRL) (ac EAT) (th CHOCOLATE)} |
| D6 | girl chases cat | {(ag GIRL) (ac CHASE) (th CAT)} {(ag GIRL) (ac RUN)} {(ag CAT) (ac FLEE) (th GIRL)} |

### Word Order Learning Curves

Fig. 2 demonstrates the acquisition of word order in M-WO, evaluated in different ambiguous contexts. We used the most likely role fillers for "subject", "object", and "verb" which correspond to "agent", "theme", and "action", to select the appropriate thematic role bigrams corresponding to syntactic position bigrams (SV,SO,OS,OV,VO,VS). Then, for each word order such as VSO, we used the product of the appropriate bigram probabilities (VS and SO) to report the probability of the word order. As can be seen in Fig. 2, word order acquisition in M-WO significantly favors the correct word order which was used to generate the data over all other possibilities. Word order acquisition in all ambiguous contexts starts moving towards the correct word order pretty quickly, while more ambiguous contexts (D2 and D3) seem to have a slower improvement rate, as they hit a plateau around the 100th situation.

### Word Learning Curves

To evaluate the incremental performance of the model, we use *mean word acquisition score* $P(object|word)$ (Alishahi and Fazly 2010) over all the word-referent mappings in the gold-standard lexicon used for data generation. Fig. 3 demonstrates the incremental improvement of mean acquisition scores for all the words (nouns and verbs), verbs and the score differences due to learning word order. As can be




(a) D1 (b) D2 (c) D3

Figure 2: Word order learning results, running the model on different datasets.

seen, the mean acquisition score improves upon receiving more data and all learning curves converge, which shows the stability of the learning algorithm. The graphs of acquisition score difference depict three different phases of the effect of joint acquisition on word learning results. First, in early trials where the context ambiguity is high as no word, object, or action is known yet and the acquired word order is still imperfect, acquisition score differences due to word order learning are mostly negative, indicating a disadvantage due to word order learning. This is intuitive as M-WO has two sources of noise (context and word order), but M-B has only one source of noise (context) in this phase. The second phase, starts when a moderate amount of data is received which facilitates improvement in both the acquired cross-situational information (context) and the acquired word order. In this phase, cross-situational information alone is not yet sufficient to converge on the real world statistics and the integration of the improved word order knowledge results in better word learning results in M-WO compared to M-B. The third phase takes places when cross-situational information inherent in data alone is sufficient to converge on real-world statistics; hence, there would be little or no score difference due to integration of even more improved word order knowledge. Furthermore, verb score differences seem to be higher than word (inclusive of verbs) score differences. It might be an artifact of having fewer verbs (10) compared to other words (30) and the averaging effect.

### One-Shot Learning

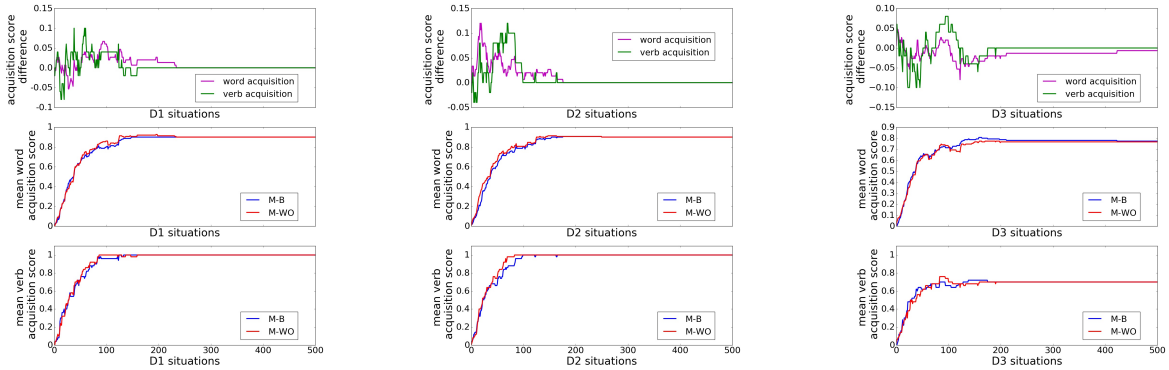
We first trained M-B and M-WO with D1 and then presented these models with the test data in D4, D5 and D6. Fig. 4 indicates better one-shot learning results for M-WO compared to M-B across different ambiguous contexts. Note that the reported acquisition scores in Fig. 4 are averaged over all the words in the training and test data (as the meaning of the shared words between training and test data are subject to change during test), which is accountable for the small score differences in one-shot learning.

### Inferring Intention in Noisy Visual Contexts

Fig. 5a depicts two phases of the effect of joint acquisition of word order on inferring intentions in ambiguous visual contexts (D3). The first phase starts with better results from M-B, which has only one source of noise (cross-situational information) compared to M-WO with two sources of noise (cross-situational information and word order knowledge). This phase is followed by the second phase during which both cross-situational information and the acquired word order are improved as a result of which M-WO results can catch up with M-B results or get even better. This phase can be followed by another phase where there would be little or no difference between the results of M-B and M-WO after receiving sufficient data on the usage of the words in the train data. Fig. 5b depicts the late effects of joint acquisition of word order on inferring intentions in ambiguous visual contexts (D6), after M-B and M-WO are first presented with train data (D1). As can be seen, M-WO exhibits higher mean accuracy in inferring the target intentions across D6 situations compared to M-B. This demonstrates the advantage of learning word order in inferring the intentions of the speaker in ambiguous contexts.

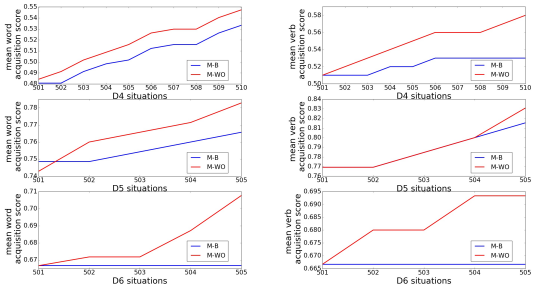
## Discussion and Conclusion

We proposed a probabilistic framework in which the knowledge of word order and word referent can be jointly learned in the absence of any prior syntactic knowledge (e.g., "subjecthood" or lexical categories). The main thesis of our framework is that transitional probabilities of the thematic roles associated with the words referring to event participants (concrete objects) and events (actions) can guide early acquisition of the notion of word order before syntactic concepts are available to the learner. Our model learns the meaning of verbs (unlike (Frank, Goodman, and Tenenbaum 2009; Sadeghi, Scheutz, and Krause 2017)) in addition to nouns and allows for the addition of synonyms. Naturalistic corpus evaluations were impossible due to the limitations of available corpus annotations (we need corpus annotations in which each scene is coded as a list of actions/events). Hence, we evaluated our model using synthetic data, varying the source and level of ambiguity in the data which includes variable-length utterances consisting of function words, prepositions, adjectives as well as nouns and action verbs paired with ambiguous (distracting events in addition to the target event) or unambiguous scenes. We used an incremental and memory-limited learning algorithm which accounts for real-world computational constraints and thus allows for implementations in online learning settings (e.g., on a robot). Fig. 2 and Fig. 3 demonstrate the stability and convergence of the learning algorithm in addition to the successful acquisition of the target word order and word refer-



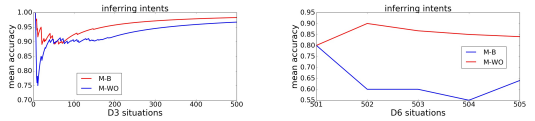

(a) D1 (b) D2 (c) D3

Figure 3: Word learning results, running the model on different datasets.




(a) word acquisition (b) verb acquisition

Figure 4: One-shot learning results using D4 (first row), D5 (second row), and D6 (third row) as the test data.




(a) averaged over D3 trials. (b) averaged over D6 trials.

Figure 5: Inferring the target event meant by the speaker.

ents. Our one-shot learning results in Fig. 4 demonstrate the advantage of joint acquisition of word order for facilitating one-shot learning (the difference between M-WO and M-B scores). Note that we used three different test datasets and although changing the dataset changed the magnitude of score differences, the differences and the direction of differences persisted which demonstrates the robustness of the reported results. Similarly, Fig. 5 demonstrates the advantage of joint acquisition of word order for inferring the intention of the speaker in ambiguous contexts where the model allows for identifying not only which action is being talked about by the speaker, but also (1) identifying from the perspective of which event participant the action is being described (e.g., "take" or "give"), and (2) how many event participants are in focus (e.g., an intransitive verb "fall" and a transitive verb "drop" differ in the number of arguments but can both be used to describe the same event using sentences such as "the box fell" or "dad dropped the box").

Our results suggest that relying on cross-situational information alone for word learning can be problematic in the presence of (1) inconsistent word-referent co-occurrence (e.g., when the perceptual referents are absent in the scene or when data contains alternative labels), and (2) ambiguity in identifying the target event (as event boundaries are not perfect and even if they were, still there are multiple ways to describe the same event using different verbs and providing graded levels of details about the event). Therefore, relying on cross-situational information alone would mean dealing with lots of ambiguity on many dimensions. Learning another source of information about language such as word order, despite adding an additional source of noise to the process of word learning in the beginning, serves to disambiguate some of that ambiguity and speed up (one-shot learning) word learning later on.

Our findings regarding the general utility of joint acquisition of word order in improving word learning results is aligned with previous computational results (Alishahi and Fazly 2010; Abend et al. 2017). However, our results differ from previous work in that they suggest that there is a time lag for the emergence of the advantage of word order learning in improving word learning results, inferring intentions, and facilitating one-shot learning, during which the acquired word order knowledge is being improved.

In future work, our framework can be extended to accommodate learning the structural rules of NPs, by adding another syntactic component such as $\Theta_{NP}$ to capture the relative order of the type of modifiers used in NPs (e.g., to learn that color modifiers cannot be followed by size modifiers but the opposite is likely as in "the large red box"). In addition to that, computational experiments with different set of thematic roles, varying the specificity versus generality of the roles, can shed light on whether adult-like notions of thematic roles are required for word order acquisition.

## Acknowledgments

This work was funded in part by Vienna Science and Technology Fund project ICT15-045.

## References


Abend, O.; Kwiatkowski, T.; Smith, N. J.; Goldwater, S.; and Steedman, M. 2017. Bootstrapping language acquisition. *Cognition* 164:116–143.

Alishahi, A., and Chrupała, G. 2012. Concurrent acquisition of word meaning and lexical categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 643–654. Association for Computational Linguistics.

Alishahi, A., and Fazly, A. 2010. Integrating syntactic knowledge into a model of cross-situational word learning. In *Proc. of CogSci*, volume 10.

Börschinger, B., and Johnson, M. 2011. A particle filter algorithm for bayesian word segmentation. In *Proceedings of the Australasian Language Technology Association Workshop*, 10–18. D. Mollá & D. Martinez.

Börschinger, B., and Johnson, M. 2012. Using rejuvenation to improve particle filtering for bayesian word segmentation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 85–89. Association for Computational Linguistics.

Brown, R., and Bellugi, U. 1964. Three processes in the child's acquisition of syntax. *Harvard educational review* 34(2):133–151.

Brown, R. 1973. *A first language: The early stages.* Harvard U. Press.

Fazly, A.; Alishahi, A.; and Stevenson, S. 2010. A probabilistic computational model of cross-situational word learning. *Cognitive Science* 34(6):1017–1063.

Frank, M. C.; Goodman, N. D.; and Tenenbaum, J. B. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* 20:578–585.

Liang, P.; Jordan, M. I.; and Klein, D. 2009. Probabilistic grammars and hierarchical dirichlet processes. *The handbook of applied Bayesian analysis*.

MacWhinney, B. 2000. The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database.

Maurits, L.; Perfors, A. F.; and Navarro, D. J. 2009. Joint acquisition of word order and word reference. Cognitive Science Society.

Medina, T. N.; Snedeker, J.; Trueswell, J. C.; and Gleitman, L. R. 2011. How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences* 108(22):9014–9019.

Pearl, L.; Goldwater, S.; and Steyvers, M. 2010. Online learning mechanisms for bayesian models of word segmentation. *Research on Language & Computation* 8(2):107–132.

Sadeghi, S., and Scheutz, M. 2017. Joint acquisition of word order and word referent in a memory-limited and incremental learner. In *Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*.

Sadeghi, S.; Scheutz, M.; and Krause, E. 2017. An embodied incremental bayesian model of cross-situational word learning. In *Proceedings of the 2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (icdl-epirob)*.

Yu, C. 2006. Learning syntax–semantics mappings to bootstrap word learning. In *Proceedings of the 28th annual conference of the cognitive science society*, volume 36.